DOCUMENT RESUME

ED 132 212 TM 005 949

AUTHOR Britan, Gerald M.
TITLE Anthropology and Evaluation: Qualitative and Quantitative Methodology in the Assessment of Action Programs.
PUB DATE [Dec 75]
NOTE 18p.; Paper presented at the Annual Meeting of the American Anthropological Association (San Francisco, California, December 1975)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Anthropology; Compensatory Education Programs; *Evaluation Methods; Formative Evaluation; *Program Evaluation; *Research Methodology; *Social Science Research; Summative Evaluation
IDENTIFIERS Ethnography

ABSTRACT

Until recently, anthropologists have had little involvement in the evaluation of social welfare and social change programs, despite the fact that such research involves not only the quantitative assessment of program results, but also analysis of informal structures and processes through which goals are formulated and results achieved. This paper considers the relevance of anthropological method and theory to such research. A detailed examination of the author's evaluation of a federal program demonstrates that while problems exist and new middle-level theory must be developed, anthropological perspectives do provide important insights and understandings. Future prospects for such anthropological research look exceedingly bright. (Author)

ANTHROPOLOGY AND EVALUATION:

Qualitative and Quantitative Methodology in the Assessment of Action Programs

presented at the annual meeting of the
American Anthropological Association,
San Francisco, December 1975.

Gerald M. Britan
Department of Anthropology
Northwestern University.

American anthropology is in a crisis. Althrough more anthropologists are being trained than ever before (about 7000 currently enrolled graduate students), traditional employment markets are shrinking. In response, the field has turned towards new possibilities for applied research, but is discovering that anthropological theory often has little problem-solving relevance. This need not be so. To cumulatively advance, our insights must be tested in a pragmatic arena, where they can be falsified, and if necessary, discarded. The evaluation of action programs provides one such setting. Not only are the techniques of anthropology essential in proper program evaluations; experience in evaluation research will add significantly to the growth of anthropological thought.

THE NATURE OF EVALUATION

During the 1960's, American society developed a deep-rooted faith in "action programs" as a solution to social problems (Williams & Evans 1969). A wide range of programs began, aimed at redistributing power and funds to uplift the disadvantaged through education, economics, public health, community development, improved ethnic relations, and so on. As the levels of commitment grew, hundreds of millions of dollars were spent on HEAD START, MODEL CITIES, JOB CORPS, and the like. By the end of the decade, however, the federal government, which funded most action programs, demanded better information about the return it was getting on its investments, and interest in evaluation research grew.

At this basic level, evaluation research attempts to determine the degree to which a program has its intended effects. Does compensatory education improve the cognitive abilities of children? Does mental rehabilitation improve psychological adjustment? Do new treatments increase the rate of drug

addiction cures? Yet, as we shall see, this summative evaluation is not so easy as it first appears.

To be more useful, both to administrators and society, evaluation should tell us not only what occurred, but how and why these results were obtained. Such process evaluation assesses program design and implementation. By elucidating mechanisms through which effects are achieved, we can directly verify the theoretical justification for a treatment, and identify possible confounding variables within the implementation process. Although process evaluation faces even greater methodological problems than the analysis of results, it is vital for an understanding of summative data.

Formative evaluation not only assesses program implementation, but applies the analysis of process and results as recommendations for improvements in program structure. While formative evaluations usually incorporate an analysis of process, they are even more of an art, without a rigorous inferential grounding. Such evaluations are usually concerned with the establishment of a new program, but could be incorporated in program design as a continuing aspect. Indeed, formative evaluation is often the responsibility of program management, and is closely identified with internal adminstrative goals. Formative consultants are "pro lem solvers", who may not strive to maintain the kind of academic dis-interest that is more characteristic of summative and process evaluation.

The distinction between forms of program evaluation are not hard and fast, nor should they be. Studies of design, process and results, all add to our understanding of what action programs accomplish and why. However, summative evaluation, despite limited rigorous application, is the dominant methodological perspective. The present paper argues that the analysis of results alone is insufficient; qualitative process evaluation poses analytical

problems of no greater magnitude and is essential for meaningful interpretation.

QUANTITATIVE EVALUATIONS OF PROGRAM RESULTS

On first glance, summative program evaluation seems very simple. A measurably disadvantaged group is provided with some ameliorative treatment, and if its performance has improved upon retesting, then the treatment is considered successful. Yet, how can we tell whether observed changes are due to the program rather than extraneous factors. An improvement in retest scores might be due to the maturation of subjects, to the lessons learned in the pre-test experience, or even to a statistical artifact of the test procedure itself.

Such problems are sometimes very difficult to see intuitively, but are crucial for our assessment of program results. A somewhat simplified example is provided by the federally funded compensatory education programs that are being implemented all over the country. Admission to remedial "treatment" is limited to those students who fall below a minimum cut-off score on a cognitive skills pretest. After a year of program participation, students are tested again and n , always a substantial improvement in scores has occurred. These stud are released, and a new crop who fell below a pre-test cut-off are admitted.

This kind of "evaluation" has been used to validate compensatory education programs and the search for increased funds all over the country. Yet, even if we were to accept the testing procedures as adequate, and assume that maturation, pre-test experience and Hawthorne effects are irrelevent, conclusions are still suspect. There are statistical artifacts which prevent such an "evaluation" from demonstrating the success of remedial treatment in im-

proving cognitive abilities.

The method of program admission-- on the basis of lowest pre-test scores-- can greatly affect program results. Such admission procedures assume that the resutls of the single pre-test provide a valid measure of the cognitive ability of students, but in reality a student's score on a given test can vary widely due to a range of chance factors. This variation is the normal distribution of pre-test scores around a mean. By taking only those students with the lowest scores into a program, we are emphasizing the downward chance variation in test scores. If variation was due to chance, on immediate re-test, the group would not repeat original scores, but duplicate the orignal population distribution (see diagram 1). In other words, regression to the population mean would lead to a substantial improvement in post-test scores even if compensatory education had no impact at all.

Comparative evaluations have been designed to deal with this problem, but there are still important obstacles to successful inference. The well-known Westinghouse/Ohio University evaluation of HEAD START provides a good example, though one in which statistical artifacts lower rather than raise our estimation of program results. The researchers were asked to design an ex-post facto study study several years after the program had begun. They proceeded by matching HEAD START participants with outsiders on the basis of cognitive pre-tests on a series of performance scales, socio-economic backgrounds and subsequent educational experience. They then compared scores on a cognitive ability post-test to see if HEAD START participants showed relative improvement. The evaluation seemed to clearly indicate that HEAD START students did no better than their "untreated" partner, and the effects of a program costing hundreds of millions of dollars were questioned (Cicirelli, et al. 1969).

These results, however, are suspect, because of the difficulties in matching pairs (Campbell & Erlebacher 1970). HEAD START programs are designed for the most disadvantaged, and nearly all eligible participants are enrolled. Thus, head start participants can be expected to score lower on a cognitive abilities test than the population at large. The matched pairs, however, would tend to be those members of the general population who had happened to score poorly on the particular pre-test used. The mean score of the population from which these individuals came would be significantly higher than the mean for disadvantaged students. Furthermore, in a retest, each group would regress towards its own population mean. Since the HEAD START mean is lower, head start students would tend to show less relative improvement, and the program would seem a failure (see diagram 2). Still unanswered, however, is whether even this improvement would have been shown in the absence of the HEAD START program.

Quantitative methodologists are well aware of the obstacles to proper inference-- the possibility of concluding that a treatment has an effect, when it really doesn't, and vice versa. Although the warnings are sometimes ignored, a variety of threats to the internal validity of an evaluation have been noted, such as differential maturation of comparison groups, variation in measuring instruments, differential mortality in treated and untreated groups, and so on (Cook & Campbell 1975). Depending on the fineness of distinctions, the list could be expanded to at least 15 or 20 items.

Sophisticated evaluation designs have increased our ability to distinguish real from apparent program effects (Campbell & Stanley 1966). Although true experiments in which subjects are randomly assigned to test and control groups are the best anser, quasi-experimental designs can provide meaningful results provided researchers are aware of the limitations

(Campbell 1974). Despite continuing examples of inadequate methodology, the theoretical basis for methodological sound summative evaluations exists. Yet, even so, there are still limits to the utility of quantitative assessments alone.

A quantitative evaluation of results is unable to answer many questions no matter how good its experimental or quasi-experimental design. It cannot tell us how a program is implemented, whether the results are transferable to other situations, or why the observed results occurred. Not all of the threats to evaluation validity can be met by considering quantitative results. Such analysis cannot tell us whether formal program goals actually reflect informal ends sought. It cannot tell us if there are differences in the way a program is implemented for different individuals or for different locations. It cannot show whether the program has changed over time. It can't elucidate subtle effects of implementation in participant selection, differential mortality or differential learning. It cannot validate the adequacy of testing methods or ascertain diffuse program affects. We must have a broader understanding of what a program does, before we can begin to explain why it succeeds or fails.

Summative evaluations treat action programs as if they were black boxes. They demonstrate what results have occurred, but do not elucidate equally important how and why questions. To understand these, we must open the black box to look at the process of program implementation and supplement the analysis of results. Such research can define the qualitative dimensions about which quantitative data can be gathered and provides a further grounding for theoretical inference. Moreover, the distinction between quantitative and qualitative understanding is not so great as we make it seem (see Campbell 1975).

THE EVALUATION OF PROCESS

While the methodological problems of quantitative evaluation are fairly great, obstacles to rigorous qualitative evaluation seem almost of another order of magnitude. Canons of rigor have yet to be established and qualitative assessments of process remain as much an art, as a science-- albeit, an art at which anthropologists are thought to excell. All we can do at present is indicate existing problems and suggest appropriate directions for pragmatic research. To understand process, we must understand how particular things are processed. In the context of action programs, we must learn how and why things occur as an individual progresses through "treatment". The ethnographic model of anthropology, with its qualitative assessment of a particular case provides a basis from which to start.

Ethnography achieves understandings by combining the knowledge of insiders and outsiders. The ethnographer has an external perspective which lets him see the importance in what an insider, from too great familiarity, dismisses as trivial. Still, to interpret his observations, the ethnographer needs a a broader context of comparison. In traditional fieldwork, this is provided by anthropological theory and a familiarity with similar regional cultures and similar field experiences. Even so, a consideration of the differences and similarities among a range of local sites is often useful. In program evaluation, the ethnographer should supplement his disciplinary background, with general knowledge about formal organizations, and about similar programs or settings. Lacking a contemporary program for comparison the ethnographer can at least consider sufficient time depth to permit an adequate comparative appraisal.

The mechanics of ethnographic technique are observation and participation, of which there are two aspects: On the one hand, the ethnographer as an outside

specialist tries to place observed events within his own categories of relationship, but at the same time he tries to understand their import for the participants themselves. The most interesting part of ethnography involves putting these two views-- the external subjectivity of the observer and the internal subjectivity of the native-- together. Both perspectives must be triangulated with developed theory to provide a fairly valid basis for ethnographic inference.

Still, there is no guarantee that the interpretation of any single ethnographer would be replicated through restudy by another. Unlike summative evaluation, which measures a few well-operationalized variables, the study of process is concerned with patterns among a much larger range of factors. The methodological problems of process analysis cannot, at present, be rigorously solved (see Campbell 1974).

Yet quantification, as such, means very little. It is often the qualitative dimensions distinguished by ethnography that provide the appropriate basis for quantitative scales. Quantitative assessments, moreover, can tell us little about <u>how</u> and <u>why</u> particular relationships exist. The import of any quantitative analysis of results rests on an independent appraisal of the causal relations involved. Ethnography tries to comprehend these relationships in their entirety, and though the problem of rigor is severe, the problem of artificiality is nearly eliminated. The final purpose of evaluation is pragmatic-- the improvement of program results. An ethnography of process provides reasonable insights about how such improvements could be made in a way that summative evaluation alone cannot.

PROCESS EVALUATION IN PRACTICE

The best way to illustrate the nature of process evaluation and its problems is through an example such as my ongoing assessment of the Experimental Technology Incentives Program (ETIP) of the National Bureau of Standards (NBS). ETIP is not a typical candidate for social science evaluation. It formal mandate, which derives from the President's 1972 Science and Technology message, is to "facilitate technological change." ETIP seeks to achieve this goal indirectly, by developing experimental poiicy changes in co-operating governmental organizations which indirectly "create an environment conducive to innovation." Despite the lack of traditional "clients" or "treatments," an assessment of ETIP poses the same questions as any study of process.

Last spring, I was asked by the National Academy of Sciences (NAS) to conduct an 18 month evaluation of the ETIP program. The questions at issue were not ETIP's final effects on technological change-- these could be considered later by technologists-- but rather an assessment of ETIP as a whole, as an experiment in organizational form within the federal bureaucracy.

Several basic questions were raised by NAS: Can an organization like ETIP actually convince federal agencies to experiment with operating policies? To what extent has ETIP been responsible for any policy changes that have occurred? Do ETIP's policy experiments reflect mandated program goals, or the internal needs of co-operating agencies? What factors affect ETIP's success in developing and implementing experiments? Do ETIP's experiments have any civilian sector effects? What is ETIP's role in the

federal environment? What modifications would improve ETIP's ability to fulfill its mission? All of these issues concern the effect of an organization, and its formal and informal operating processes, on other organizations within the same over-arching environment. Although anthropologists have avoided studies of bureaucracies and administrative elites (despite suggestions that the vacuum be filled (e.g. Foster 1969)), these issues are important to the discipline.

Their resolution, however, is difficult, and must certainly transcend any straight-forward summation of results. Some of the problems involved in this kind of evaluation should be mentioned: It is difficult to conduct a real-time appraisal of ETIP, since it is a dynamically evolving program whose goals, clients, projects and personnel change during the course of study. Because the program is still developing there is a lack of data on its civilian sector effects, and summative evaluations of particular FTIP projects will not be available until late in 1977. Furthermore, it is very hard to measure many of the ephemeral effects, such as changes in agency policy, which are the direct concern of the study. Finally, there is no baseline data-- similar programs which could be used for comparison. In general, the possibilities for truely rigorous inference are limited. Still, since the research provided an opportunity to study American social and administrative organization at a level that is rarely possible, I decided to go ahead.

At the time of my introduction to ETIP, the program had a staff of 12 and was conducting more than 100 projects with over 40 federal agency and private clients. Materials on a single project sometimes filled an entire file drawer. Often 4 or 5 co-operating offices were involved,

and more than a score of individuals. Faced with access to so much data, I first had to sit back and develop a strategy with which to proceed. My background in anthropological ethnography provided, I think, the best starting point.

The first task was to reach a better understanding about the program itself, by conducting an ethnography of ETIP which could answer a whole series of "who", "what", "where", and "how" questions. Since internal documentary evidence seemed suspiciously one-sided, I searched for as many different sources of data as possible. Only after determining precisely what ETIP did could I turn to the program's relationships with client agencies.

The statements of goals and procedures found in formal documents were supplemented by an analysis of entire stacks of bureaucratic paperwork-- memos, schedules, budgets-- which put the formal evidence in another light. A whole range of informal documents about particular ETIP projects provided a basis for quantified measures of project type and project success. Most important, however, were the interviews, which were conducted with both current and former staff members through a variety of techniques. Only when information from all these sources was combined and analyzed, could a clear idea of ETIP's operations be developed. Although many details must await further confirmation, ETIP is clearly a rather different organization in practice than it appears in formal design, with goals that are often far removed from any issues of "technological change."

This last fact raises an important problem. Although ETIP sometimes seems to go beyond its formal mandate, many such activities are both

useful and successful. What, however, should be the goals against which the programs success is measured?

The issue is difficult, for any action program involves a diversity of interests whose goals and implementation policies differ (Krause & Howard 1975). An ethnography of the implementation process, at least, can begin to show how formal goals in an action program are modified in practice. The definition of such informal goals cannot be obtained through summative evaluation alone. On the other hand, the qualitative appraisal of ETIP unearthed a number of dimensions (e.g. project type, project objectives, staff committment, etc.) for which appropriate quantitative measures were developed and analyzed.

Questions about goal orientation are even more important in the current stage of research, the study of ETIP/agency relations. The implementation of action programs is a political process (Krause & Howard 1975), and differences between ETIP and its agency clients are to be anticipated. To what extent, though, do differences in goal orientation, funcing priorities, or implementation procedures affect the success of the ETIP program?

To find out, a series of case studies are being investigated covering the range of project types discovered in the earlier ethnography. Again a multi-method approach is being used to provide as much diversity in the sources of data as possible. These in-depth case studies are expected to define the parameters for a more rapid survey of other ETIP/agency relationships and of civilian sector responses to particular ETIP projects.

In the last stage of research all sources of data will be comined, and lacunae located and filled. An analysis of both qualitative and quantitative relationships will proceed. While the final result will not be an entirely rigorous assessment of ETIP's performance, an analytical background for later interpretation will have been found.

CONCLUSIONS

Process evaluations of action programs are closely akin to anthropological ethnographies. The researcher must integrate a diverse body of data, encompassing outsider and insider points of view, in order to explicate social process. The outcome is an in-depth undersanding which enables the definition and interpretation of appropriate quantitative measures. While the objective validity of such an ethnography is difficult to measure, some level of validity does exist, for an ethnographer cannot simply manipulate variables ex post facto to substantiate his conclusions (Campbell 1974). The ethnographer is concerned with patterns, and any particular hypothesise has multiple implications which must be demonstrated. Moreover, a good ethnographer must successfully account for similarities and differences between the observers and the native's point of view.

Although ethnographic techniques are still imperfect, without them only common-sense would be available to help interpret quantitative measures. Certainly, a thorough-going ethnography provides a better guide. Experience has taught us that social systems often behave in a counter-intuitive fashion (Forrester 1968), and common-sense alone provides a limit basis for the design of action programs. Qualitative assessments of program

processes are essential in interpreting summative results and designing program improvements.

Still, even if ethnographic techniques are important to action research, why should anthropologists be involved? Mundane facts could be cited, such as the need for solutions to practical problems, or the job-shortage in academia, but evaluation research has a theoretical relevance to anthropology as well.

The focus of this symposium is the place of theory in problem oriented anthropology. Thus far, anthropological research has not yielded a cumulative growth of theory. But our discipline does not face this problem alone. Recent critics have noted the general deficiencies of social science understandings of human behavior (Gordon & Morse 1975). The major problem facing evaluation research is not inadequate methodology-- new tools can be developed-- but a lack of appropriate middle level theory. Yet evaluation research, the assessment of action programs, has an enormous potential for developing such theory. Its findings are applied as social policies, and the validity of conclusions is subject to rapid real world testing.

Problem oriented research, such as program evaluation, is crucial to anthropology. Theories can only be proved if they are applied to concrete situations where they can also be falsified. Applied anthropology is not a poor relation to the mainstream, but must lead in the development of new understandings of man's place in the world. Evaluation research provides not merely a new employment option, but an opportunity to re-integrate anthropological theory and practice as well.

DIAGRAM 1: Regression to the Mean in Compensatory Education

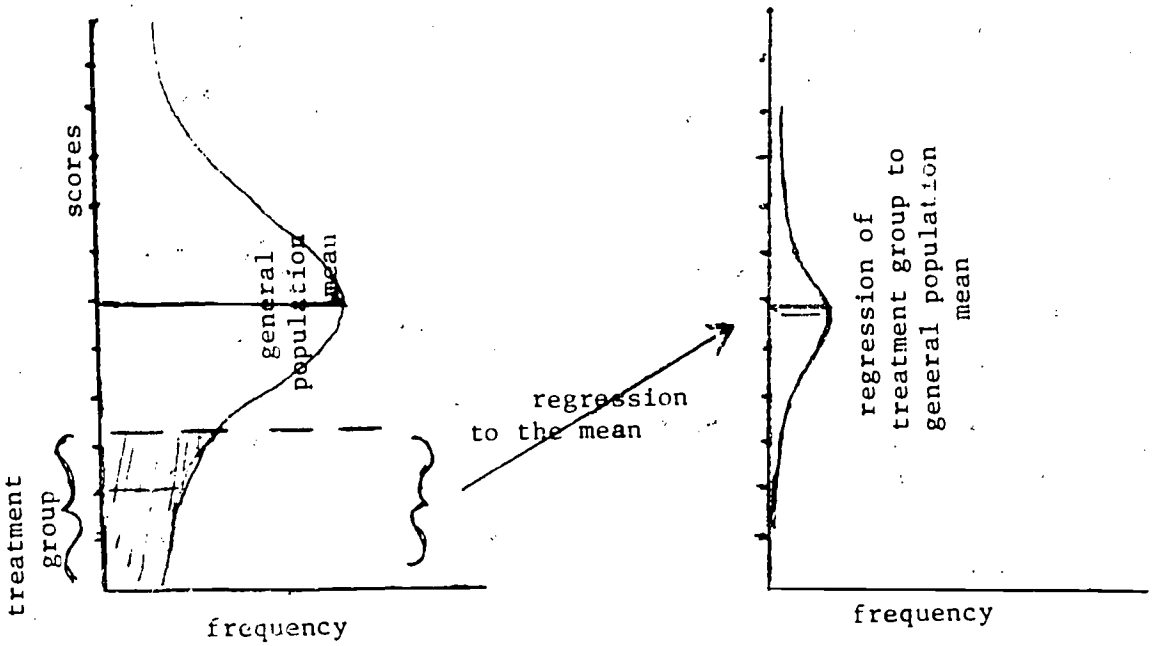
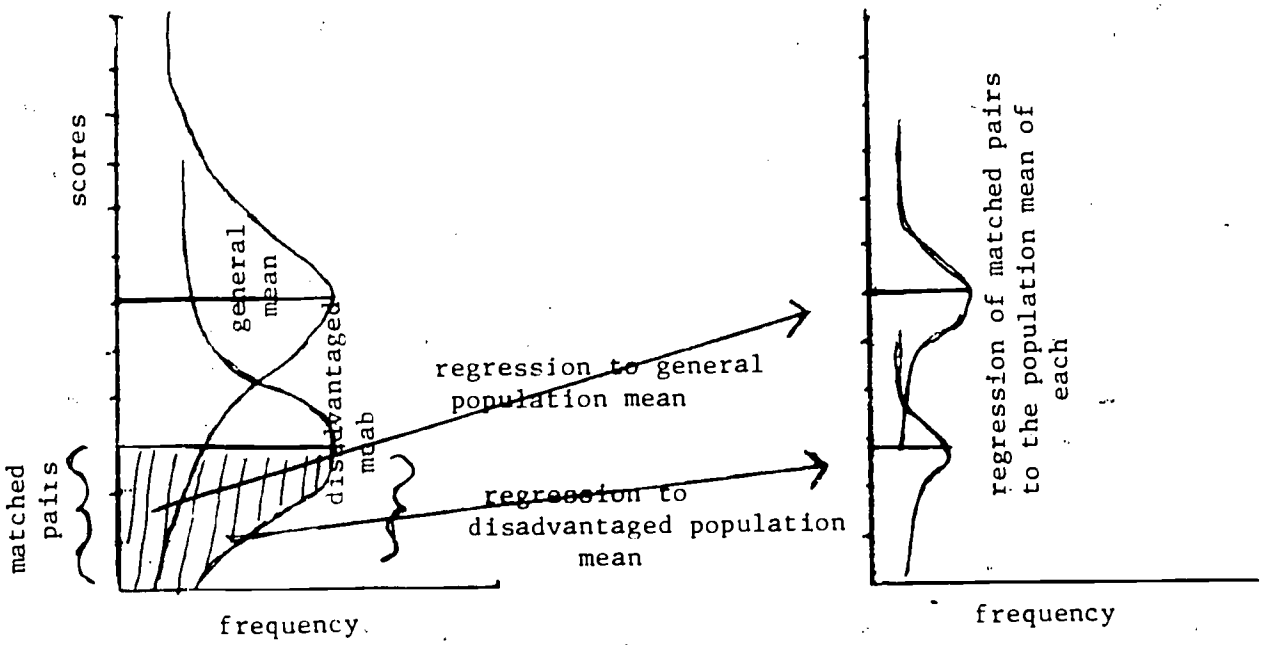


DIAGRAM 2: Regression to the Mean with Matched Pairs



REFERENCES CITED

Campbell, Donald T.
1974a "Quasi-experimental Designs" In: H.W. Riecken & R.F Boruch (eds) Social Experimentation. New York, Academic Press.

1974b "Qualitative Knowing in Action Research" forthcoming in the Journal of Social Issues.

1975 "Assessing the Impact of Planned Social Change" In: G. Lyons (ed), Social Research and Public Policies. New Hampshire, University Press of New England.

Campbell, Donald T. & Albert Erlebacher
1970 "How Regression Artifacts in Quasi-experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful" In: J. Hellmuth (ed), Compensatory Education. New York, Brunner/Mazel.

Cicirelli, V. et al.
1969 The Impact of Head Start. Washington, Dept. of Commerce.

Cook, Thomas D. & Donald T. Campbell
1975 "The Design and Conduct of Quasi-Experiments and True Experiments in Field Settings" In M.D. Hunnette (ed), Handbook of Industrial and Organizational Research. N.Y., Rand McNally.

Forrester, Jay
1968 "Counter-intuitive Behavior in Social Systems"

Gordon, Gerald & Edward V. Morse
1975 "Evaluation Research" In: The Annual Review of Sociology I:

Krause, Merton S. & Kenneth I. Howard
1975 "Program Evaluation in the Public Interest" In: Readings in Program Evaluation Research. Evanston, Northwestern University.

Williams, Walter & John W. Evans
1969 "The Politics of Evaluation" Annals of the American Academy of Political and Social Science 385: 118-132.

proving cognitive abilities.

The method of program admission-- on the basis of lowest pre-test scores-- can greatly affect program results. Such admission procedures assume that the resutls of the single pre-test provide a valid measure of the cognitive ability of students, but in reality a student's score on a given test can vary widely due to a range of chance factors. This variation is the normal distribution of pre-test scores around a mean. By taking only those students with the lowest scores into a program, we are emphasizing the downward chance variation in test scores. If variation was due to chance, on immediate re-test, the group would not repeat original scores, but duplicate the orignal population distribution (see diagram 1). In other words, regression to the population mean would lead to a substantial improvement in post-test scores even if compensatory education had no impact at all.

Comparative evaluations have been designed to deal with this problem, but there are still important obstacles to successful inference. The well-known Westinghouse/Ohio University evaluation of HEAD START provides a good example, though one in which statistical artifacts lower rather than raise our estimation of program results. The researchers were asked to design an ex-post facto study study several years after the program had begun. They proceeded by matching HEAD START participants with outsiders on the basis of cognitive pre-tests on a series of performance scales, socio-economic backgrounds and subsequent educational experience. They then compared scores on a cognitive ability post-test to see if HEAD START participants showed relative improvement. The evaluation seemed to clearly indicate that HEAD START students did no better than their "untreated" partner, and the effects of a program costing hundreds of millions of dollars were questioned (Cicirelli, et al. 1969).

These results, however, are suspect, because of the difficulties in matching pairs (Campbell & Erlebacher 1970). HEAD START programs are designed for the most disadvantaged, and nearly all eligible participants are enrolled. Thus, head start participants can be expected to score lower on a cognitive abilities test than the population at large. The matched pairs, however, would tend to be those members of the general population who had happened to score poorly on the particular pre-test used. The mean score of the population from which these individuals came would be significantly higher than the mean for disadvantaged students. Furthermore, in a retest, each group would regress towards its own population mean. Since the HEAD START mean is lower, head start students would tend to show less relative improvement, and the program would seem a failure (see diagram 2). Still unanswered, however, is whether even this improvement would have been shown in the absence of the HEAD START program.

Quantitative methodologists are well aware of the obstacles to proper inference-- the possibility of concluding that a treatment has an effect, when it really doesn't, and vice versa. Although the warnings are sometimes ignored, a variety of threats to the internal validity of an evaluation have been noted, such as differential maturation of comparison groups, variation in measuring instruments, differential mortality in treated and untreated groups, and so on (Cook & Campbell 1975). Depending on the fineness of distinctions, the list could be expanded to at least 15 or 20 items.

Sophisticated evaluation designs have increased our ability to distinguish real from apparent program effects (Campbell & Stanley 1966). Although true experiments in which subjects are randomly assigned to test and control groups are the best anser, quasi-experimental designs can provide meaningful results provided researchers are aware of the limitations

(Campbell 1974). Despite continuing examples of inadequate methodology, the theoretical basis for methodological sound summative evaluations exists. Yet, even so, there are still limits to the utility of quantitative assessments alone.

A quantitative evaluation of results is unable to answer many questions no matter how good its experimental or quasi-experimental design. It cannot tell us how a program is implemented, whether the results are transferable to other situations, or why the observed results occurred. Not all of the threats to evaluation validity can be met by considering quantitative results. Such analysis cannot tell us whether formal program goals actually reflect informal ends sought. It cannot tell us if there are differences in the way a program is implemented for different individuals or for different locations. It cannot show whether the program has changed over time. It can't elucidate subtle effects of implementation in participant selection, differential mortality or differential learning. It cannot validate the adequacy of testing methods or ascertain diffuse program affects. We must have a broader understanding of what a program does, before we can begin to explain why it succeeds or fails.

Summative evaluations treat action programs as if they were black boxes. They demonstrate what results have occurred, but do not elucidate equally important how and why questions. To understand these, we must open the black box to look at the process of program implementation and supplement the analysis of results. Such research can define the qualitative dimensions about which quantitative data can be gathered and provides a further grounding for theoretical inference. Moreover, the distinction between quantitative and qualitative understanding is not so great as we make it seem (see Campbell 1975).

THE EVALUATION OF PROCESS

While the methodological problems of quantitative evaluation are fairly great, obstacles to rigorous qualitative evaluation seem almost of another order of magnitude. Canons of rigor have yet to be established and qualitative assessments of process remain as much an art, as a science-- albeit, an art at which anthropologists are thought to excell. All we can do at present is indicate existing problems and suggest appropriate directions for pragmatic research. To understand process, we must understand how particular things are processed. In the context of action programs, we must learn how and why things occur as an individual progresses through "treatment". The ethnographic model of anthropology, with its qualitative assessment of a particular case provides a basis from which to start.

Ethnography achieves understandings by combining the knowledge of insiders and outsiders. The ethnographer has an external perspective which lets him see the importance in what an insider, from too great familiarity, dismisses as trivial. Still, to interpret his observations, the ethnographer needs a a broader context of comparison. In traditional fieldwork, this is provided by anthropological theory and a familiarity with similar regional cultures and similar field experiences. Even so, a consideration of the differences and similarities among a range of local sites is often useful. In program evaluation, the ethnographer should supplement his disciplinary background, with general knowledge about formal organizations, and about similar programs or settings. Lacking a contemporary program for comparison the ethnographer can at least consider sufficient time depth to permit an adequate comparative appraisal.

The mechanics of ethnographic technique are observation and participation, of which there are two aspects: On the one hand, the ethnographer as an outside

specialist tries to place observed events within his own categories of relationship, but at the same time he tries to understand their import for the participants themselves. The most interesting part of ethnography involves putting these two views-- the external subjectivity of the observer and the internal subjectivity of the native-- together. Both perspectives must be triangulated with developed theory to provide a fairly valid basis for ethnographic inference.

Still, there is no guarantee that the interpretation of any single ethnographer would be replicated through restudy by another. Unlike summative evaluation, which measures a few well-operationalized variables, the study of process is concerned with patterns among a much larger range of factors. The methodological problems of process analysis cannot, at present, be rigorously solved (see Campbell 1974).

Yet quantification, as such, means very little. It is often the qualitative dimensions distinguished by ethnography that provide the appropriate basis for quantitative scales. Quantitative assessments, moreover, can tell us little about <u>how</u> and <u>why</u> particular relationships exist. The import of any quantitative analysis of results rests on an independent appraisal of the causal relations involved. Ethnography tries to comprehend these relationships in their entirety, and though the problem of rigor is severe, the problem of artificiality is nearly eliminated. The final purpose of evaluation is pragmatic-- the improvement of program results. An ethnography of process provides reasonable insights about how such improvements could be made in a way that summative evaluation alone cannot.

PROCESS EVALUATION IN PRACTICE

The best way to illustrate the nature of process evaluation and its problems is through an example such as my ongoing assessment of the Experimental Technology Incentives Program (ETIP) of the National Bureau of Standards (NBS). ETIP is not a typical candidate for social science evaluation. It formal mandate, which derives from the President's 1972 Science and Technology message, is to "facilitate technological change." ETIP seeks to achieve this goal indirectly, by developing experimental policy changes in co-operating governmental organizations which indirectly "create an environment conducive to innovation." Despite the lack of traditional "clients" or "treatments," an assessment of ETIP poses the same questions as any study of process.

Last spring, I was asked by the National Academy of Sciences (NAS) to conduct an 18 month evaluation of the ETIP program. The questions at issue were not ETIP's final effects on technological change-- these could be considered later by technologists-- but rather an assessment of ETIP as a whole, as an experiment in organizational form within the federal bureaucracy.

Several basic questions were raised by NAS: Can an organization like ETIP actually convince federal agencies to experiment with operating policies? To what extent has ETIP been responsible for any policy changes that have occurred? Do ETIP's policy experiments reflect mandated program goals, or the internal needs of co-operating agencies? What factors affect ETIP's success in developing and implementing experiments? Do ETIP's experiments have any civilian sector effects? What is ETIP's role in the

federal environment? What modifications would improve ETIP's ability to fulfill its mission? All of these issues concern the effect of an organization, and its formal and informal operating processes, on other organizations within the same over-arching environment. Although anthropologists have avoided studies of bureaucracies and administrative elites (despite suggestions that the vacuum be filled (e.g. Foster 1969)), these issues are important to the discipline.

Their resolution, however, is difficult, and must certainly transcend any straight-forward summation of results. Some of the problems involved in this kind of evaluation should be mentioned: It is difficult to conduct a real-time appraisal of ETIP, since it is a dynamically evolving program whose goals, clients, projects and personnel change during the course of study. Because the program is still developing there is a lack of data on its civilian sector effects, and summative evaluations of particular ETIP projects will not be available until late in 1977. Furthermore, it is very hard to measure many of the ephemeral effects, such as changes in agency policy, which are the direct concern of the study. Finally, there is no baseline data-- similar programs which could be used for comparison. In general, the possibilities for truely rigorous inference are limited. Still, since the research provided an opportunity to study American social and administrative organization at a level that is rarely possible, I decided to go ahead.

At the time of my introduction to ETIP, the program had a staff of 12 and was conducting more than 100 projects with over 40 federal agency and private clients. Materials on a single project sometimes filled an entire file drawer. Often 4 or 5 co-operating offices were involved,